Machine Learning 2 - Homework Assignment 1

Alexandra Lindt

October 2, 2019

Problem 1

$$\mathbb{E}[y] = \mathbb{E}[x+z] = \mathbb{E}[x] + \mathbb{E}[z] = \mu_x + \mu_z$$

$$\begin{aligned}
Cov[y,y] &= Var[y] \\
&= \mathbb{E}[y^2] - \mathbb{E}[y]^2 \\
&= \mathbb{E}[(x+z)^2] - \mathbb{E}[x+z]^2 \\
&= \mathbb{E}[x^2 + 2xz + z^2] - (\mathbb{E}[x] + \mathbb{E}[z])^2 \\
&= \mathbb{E}[x^2] + \mathbb{E}[2xz] + \mathbb{E}[z^2] - (\mathbb{E}[x]^2 + 2\mathbb{E}[x]\mathbb{E}[z] + \mathbb{E}[z]^2) \\
&= \mathbb{E}[x^2] - \mathbb{E}[x]^2 + \mathbb{E}[z^2] - \mathbb{E}[z]^2 + 2\mathbb{E}[xz] - 2\mathbb{E}[x]\mathbb{E}[z] \\
\star &= Var(x) + Var(z) + 2\mathbb{E}[xz] - 2\mathbb{E}[xz] \\
&= Var(x) + Var(z) \\
&= \Sigma_x + \Sigma_z
\end{aligned}$$

$\star$: Since x and z are independent, it holds that $\mathbb{E}[xz] = \mathbb{E}[x]\mathbb{E}[z]$.

Problem 2

1.

$$p(\mathcal{X}|\mu,\Sigma) \overset{\text{i.i.d.}}{=} \prod_{i=1}^{N} p(x_i|\mu,\Sigma) = \prod_{i=1}^{N} \mathcal{N}(x_i|\mu,\Sigma)$$

2.

$$p(\mu|\mathcal{X},\Sigma,\mu_0,\Sigma_0) = \frac{p(\mathcal{X}|\mu,\Sigma)p(\mu|\mu_0,\Sigma_0)}{p(\mathcal{X})} = \frac{\mathcal{N}(\mathcal{X}|\mu,\Sigma)\mathcal{N}(\mu|\mu_0,\Sigma_0)}{\int_{\mu_i} \mathcal{N}(\mathcal{X}|\mu_i,\Sigma)\mathcal{N}(\mu_i|\mu_0,\Sigma_0)d\mu_i}$$

3.

$$
\begin{aligned}
p(\mu|\mathcal{X},\Sigma,\mu_0,\Sigma_0) &= \frac{\mathcal{N}(\mathcal{X}|\mu,\Sigma)\mathcal{N}(\mu|\mu_0,\Sigma_0)}{\int_{\mu_i}\mathcal{N}(\mathcal{X}|\mu_i,\Sigma)\mathcal{N}(\mu_i|\mu_0,\Sigma_0)d\mu_i} \\
&\propto \mathcal{N}(\mathcal{X}|\mu,\Sigma)\mathcal{N}(\mu|\mu_0,\Sigma_0) \\
&= \mathcal{N}(\mu|\mu_0,\Sigma_0)\prod_{i=1}^{N}\mathcal{N}(x_i|\mu,\Sigma) \\
&= \frac{1}{(2\pi)^{\frac{D}{2}}|\Sigma_0|^{\frac{1}{2}}}\exp\Big(-\frac{1}{2}(\mu-\mu_0)^T\Sigma_0^{-1}(\mu-\mu_0)\Big) \\
&\quad\cdot\frac{1}{(2\pi)^{\frac{DN}{2}}|\Sigma|^{\frac{N}{2}}}\exp\Big(\sum_{i=1}^{N}-\frac{1}{2}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)\Big) \\
&\propto \exp\Big(-\frac{1}{2}(\mu-\mu_0)^T\Sigma_0^{-1}(\mu-\mu_0)+\sum_{i=1}^{N}-\frac{1}{2}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)\Big) \\
&= \exp\Big(-\frac{1}{2}\Big(\mu^T\Sigma_0^{-1}\mu-2\mu^T\Sigma_0^{-1}\mu_0+\mu_0^T\Sigma_0^{-1}\mu_0+\sum_{i=1}^{N}x_i^T\Sigma^{-1}x_i-2\mu^T\Sigma^{-1}x_i+\mu^T\Sigma^{-1}\mu\Big)\Big) \\
&\propto \exp\Big(-\frac{1}{2}\Big(\mu^T(\Sigma_0^{-1}+N\Sigma^{-1})\mu-2\mu^T\Sigma_0^{-1}\mu_0-\sum_{i=1}^{N}2\mu^T\Sigma^{-1}x_i\Big)\Big) \\
&= \exp\Big(-\frac{1}{2}\Big(\mu^T(\Sigma_0^{-1}+N\Sigma^{-1})\mu-2\mu^T(\Sigma_0^{-1}\mu_0+\Sigma^{-1}\sum_{i=1}^{N}x_i)\Big)\Big) \\
\star &= \exp\Big(-\frac{1}{2}\Big(\mu^T\Sigma_N^{-1}\mu-2\mu^T(\Sigma_0^{-1}\mu_0+\Sigma^{-1}\sum_{i=1}^{N}x_i)\Big)\Big) \\
&= \exp\Big(-\frac{1}{2}\Big(\mu^T\Sigma_N^{-1}\mu-2\mu^T\Sigma_N^{-1}\Sigma_N(\Sigma_0^{-1}\mu_0+\Sigma^{-1}\sum_{i=1}^{N}x_i)\Big)\Big) \\
\star &= \exp\Big(-\frac{1}{2}\Big(\mu^T\Sigma_N^{-1}\mu-2\mu^T\Sigma_N^{-1}\mu_n\Big)\Big) \\
&\propto \exp\Big(-\frac{1}{2}\Big(\mu^T\Sigma_N^{-1}\mu-2\mu^T\Sigma_N^{-1}\mu_n+\mu_n^T\Sigma_N^{-1}\mu_n\Big)\Big) \\
&= \exp\Big(-\frac{1}{2}(\mu-\mu_N)^T\Sigma_N^{-1}(\mu-\mu_N)\Big) \\
&\propto \mathcal{N}(\mu|\mu_n,\Sigma_n)
\end{aligned}
$$

$\star$: In this equation, we have defined Σ_N and μ_N as

$$
\begin{aligned}
\Sigma_N &= (\Sigma_0^{-1}+N\Sigma^{-1})^{-1} \\
\mu_N &= \Sigma_N(\Sigma_0^{-1}\mu_0+\Sigma^{-1}\sum_{i=1}^{N}x_i)
\end{aligned}
$$

4.

$$
\begin{aligned}
\frac{\partial}{\partial \mu} \log p(\mu|\mathcal{X}, \Sigma, \mu_0, \Sigma_0) &= 0 \\
\frac{\partial}{\partial \mu} \log \left(\frac{\mathcal{N}(\mathcal{X}|\mu, \Sigma)\mathcal{N}(\mu|\mu_0, \Sigma_0)}{\int_{\mu_i} \mathcal{N}(\mathcal{X}|\mu_i, \Sigma)\mathcal{N}(\mu_i|\mu_0, \Sigma_0) d\mu_i} \right) &= 0 \\
\frac{\partial}{\partial \mu} \big(\log(\mathcal{N}(\mathcal{X}|\mu, \Sigma)) + \log(\mathcal{N}(\mu|\mu_0, \Sigma_0)) \big) &= 0 \\
\frac{\partial}{\partial \mu} \Big(-\frac{1}{2}(\mu - \mu_0)^T \Sigma_0^{-1} (\mu - \mu_0) + \sum_{i=1}^{N} -\frac{1}{2}(x_i - \mu)^T \Sigma^{-1} (x_i - \mu) \Big) &= 0 \\
\star \quad 2\Sigma_0^{-1}(\mu - \mu_0) + \sum_{i=1}^{N} -2\Sigma^{-1}(x_i)\mu) &= 0 \\
\Sigma_0^{-1}\mu - \Sigma_0^{-1}\mu_0 - \Sigma^{-1} \sum_{i=1}^{N} x_i + N\Sigma^{-1}\mu &= 0 \\
(\Sigma_0^{-1} + N\Sigma^{-1})\mu &= \Sigma_0^{-1}\mu_0 + \Sigma^{-1} \sum_{i=1}^{N} x_i \\
\Sigma_N^{-1}\mu &= \Sigma_0^{-1}\mu_0 + \Sigma^{-1} \sum_{i=1}^{N} x_i \\
\mu &= \Sigma_N \Big(\Sigma_0^{-1}\mu_0 + \Sigma^{-1} \sum_{i=1}^{N} x_i \Big) \\
\implies \mu_{MAP} &= \mu_N
\end{aligned}
$$

$\star$: Since Σ and Σ_0 are symmetric, we can use the following derivatives given in formula 85 and 86 of the matrix cookbook:

$$
\begin{aligned}
\frac{\partial}{\partial x}(x - s)^T W (x - s) &= 2W(x - s) \\
\frac{\partial}{\partial s}(x - s)^T W (x - s) &= -2W(x - s)
\end{aligned}
$$

Problem 3

1.

Using the maximum likelihood estimator given in equation 2.7 in the Bishop, we obtain:

$$
\mu_{MLE} = \frac{1}{N} \sum_{i=1}^{N} x_i = \frac{1}{3}(1 + 1 + 1) = 1
$$

As explained in the Bishop, this is an example of over-fitting associated with maximum likelihood estimation and means we would assume every future coin toss to come up with head.

2.

Using the given prior, we can find the maximum a posteriori solution μ_{MAP} by finding the maximum of the log posterior:

$$
\begin{aligned}
\frac{\partial}{\partial \mu} \log p(\mu|\mathcal{D}, a, b) &= 0 \\
\frac{\partial}{\partial \mu} \log(p(\mathcal{D}|\mu)p(\mu|a, b)) &= 0 \\
\frac{\partial}{\partial \mu} \log\Big((\prod_{n=1}^{N} \mu^{x_n}(1-\mu)^{1-x_n}) \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \mu^{a-1}(1-\mu)^{b-1} \Big) &= 0 \\
\frac{\partial}{\partial \mu} \log\Big((\prod_{n=1}^{N} \mu^{x_n}(1-\mu)^{1-x_n}) \mu^{a-1}(1-\mu)^{b-1} \Big) &= 0 \\
\frac{\partial}{\partial \mu} \Big((\sum_{n=1}^{N} \log(\mu^{x_n}) + \log((1-\mu)^{1-x_n})) + \log(\mu^{a-1}) + \log((1-\mu)^{b-1}) \Big) &= 0 \\
\frac{\partial}{\partial \mu} \Big((\sum_{n=1}^{N} x_n \log(\mu)) + (\sum_{n=1}^{N} (1-x_n) \log(1-\mu)) + (a-1)\log(\mu) + (b-1)\log(1-\mu) \Big) &= 0 \\
\star \quad \frac{\partial}{\partial \mu} \Big(m\log(\mu) + l\log(1-\mu) + (a-1)\log(\mu) + (b-1)\log(1-\mu) \Big) &= 0 \\
\frac{m+(a-1)}{\mu} - \frac{l+(b-1)}{1-\mu} &= 0 \\
\frac{m+a-1}{\mu} &= \frac{l+b-1}{1-\mu} \\
(1-\mu)(m+a-1) &= \mu(l+b-1) \\
(m+a-1) &= \mu(m+a-1+l+b-1) \\
\implies \mu_{MAP} &= \frac{(m+a-1)}{(m+a+l+b-2)}
\end{aligned}
$$

$\star$: Following the notation of Bishop, we set m as the number of heads and l as the number of tails so far:

$$
m = \sum_{n=1}^{N} x_n \qquad \text{and} \qquad l = \sum_{n=1}^{N} 1 - x_n
$$

3.

$$
\begin{aligned}
\mathbb{E}(\mu|\mathcal{D}) &= \frac{m+a}{m+a+l+b} = \frac{a}{m+a+l+b} + \frac{m}{m+a+l+b} \\
&= \frac{a(a+b)}{(m+a+l+b)(a+b)} + \frac{m(m+l)}{(m+a+l+b)(m+l)} \\
&= \frac{\mathbb{E}(\mu|a,b)(a+b)}{(m+a+l+b)} + \frac{\mu_{MLE}(m+l)}{(m+a+l+b)}
\end{aligned}
$$

It is apparent that $\frac{(m+l)}{(m+a+l+b)} + \frac{(a+b)}{(m+a+l+b)} = 1$. Since $a, b > 0$ and $m, l \geq 0$ per definition, we further know that $\frac{(a+b)}{(m+a+l+b)} < 1$ and $\frac{(m+l)}{(m+a+l+b)} < 1$ and therefore that $\mathbb{E}(\mu|\mathcal{D}) \neq \mu_{MLE} \neq \mathbb{E}(\mu|a, b)$. Together, this means that $\mathbb{E}(\mu|\mathcal{D})$ has to be **between** μ_{MLE} and $\mathbb{E}(\mu|a, b)$.

Problem 4

1.

(i)

$$\text{Pois}(k|\lambda) = \frac{\lambda^k e^{-\lambda}}{k!} = \frac{1}{k!} e^{\log(\lambda^k) - \lambda} = \frac{1}{k!} e^{k \log(\lambda) - \lambda} = \frac{1}{k!} e^{-\lambda} e^{\log(\lambda) k} = h(k) g(\eta) e^{\eta^T u(k)}$$

$$\implies \qquad h(k) = \frac{1}{k!} \qquad \eta = \log(\lambda) \qquad g(\eta) = e^{-\lambda} = e^{-e^{\eta}} \qquad u(k) = k$$

(ii)

$$\begin{aligned}
\text{Gam}(\tau|a, b) &= \frac{1}{\Gamma(a)} b^a \tau^{a-1} e^{-b\tau} = \frac{1}{\Gamma(a)} b^a e^{\log(\tau^{a-1}) - b\tau} = \frac{b^a}{\Gamma(a)} e^{(a-1)\log(\tau) - b\tau} \\
&= \frac{b^a}{\Gamma(a)} \exp\left(\begin{bmatrix} a-1 \\ -b \end{bmatrix}^T \begin{bmatrix} \log(\tau) \\ \tau \end{bmatrix} \right) = h(\tau) g(\eta) e^{\eta^T u(\tau)}
\end{aligned}$$

$$\implies \qquad h(\tau) = 1 \qquad \eta = \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} = \begin{bmatrix} a-1 \\ -b \end{bmatrix} \qquad g(\eta) = \frac{b^a}{\Gamma(a)} = \frac{-\eta_2^{\eta_1+1}}{\Gamma(\eta_1+1)} \qquad u(\tau) = \begin{bmatrix} \log(\tau) \\ \tau \end{bmatrix}$$

(iii)

The Cauchy distribution has no defined mean and variance, which means it is not part of the exponential family.

(iv)

$$\begin{aligned}
\text{vonMises}(x|\kappa, \mu) &= \frac{1}{2\pi I_0(\kappa)} e^{\kappa \cos(x-\mu)} = \frac{1}{2\pi I_0(\kappa)} e^{\kappa(\cos(x)\cos(\mu) + \sin(x)\sin(\mu))} \\
&= \frac{1}{2\pi I_0(\kappa)} \exp\left(\begin{bmatrix} \kappa\cos(\mu) \\ \kappa\sin(\mu) \end{bmatrix}^T \begin{bmatrix} \cos(x) \\ \sin(x) \end{bmatrix} \right) = h(x) g(\eta) e^{\eta^T u(x)}
\end{aligned}$$

$$\implies \qquad h(x) = \frac{1}{2\pi} \qquad \eta = \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} = \begin{bmatrix} \kappa\cos(\mu) \\ \kappa\sin(\mu) \end{bmatrix} \qquad u(x) = \begin{bmatrix} \cos(x) \\ \sin(x) \end{bmatrix}$$

Solving for κ :

$$\cos(\mu)^2 + \sin(\mu)^2 = 1$$
$$(\frac{\eta_1}{\kappa})^2 + (\frac{\eta_2}{\kappa})^2 = 1$$
$${\eta_1}^2 + {\eta_2}^2 = \kappa^2$$
$$\sqrt{{\eta_1}^2 + {\eta_2}^2} = \kappa$$
$$\implies g(\eta) = \frac{1}{I_0(\kappa)} = \frac{1}{I_0(\sqrt{{\eta_1}^2 + {\eta_2}^2})}$$

2.

Using equation 2.226 from the Bishop, we obtain

(i)

$$\mathbb{E}[u(k)] = -\frac{\partial}{\partial \eta} \log(e^{-e^\eta}) = -\frac{\partial}{\partial \eta} - e^\eta = e^\eta = \lambda$$

$$\mathbb{E}[u(k)]^2 = -\frac{\partial}{\partial \eta}^2 \log(e^{-e^\eta}) = \frac{\partial}{\partial \eta}^2 e^\eta = e^\eta = \lambda$$

(ii)

$$\begin{aligned}\mathbb{E}[u_2(k)] &= -\frac{\partial}{\partial \eta_2} \log(\frac{-\eta_2^{\eta_1+1}}{\Gamma(\eta_1+1)}) = -\frac{\partial}{\partial \eta_2}((\eta_1+1)\log(-\eta_2) - \log(\Gamma(\eta_1+1))) \\ &= -\frac{\partial}{\partial \eta_2}((\eta_1+1)\log(-\eta_2)) = (\eta_1+1)\frac{1}{\eta_2} = -\frac{a}{b}\end{aligned}$$

$$\begin{aligned}\mathbb{E}[u_2(k)]^2 &= -\frac{\partial}{\partial \eta_2}^2 \log(\frac{-\eta_2^{\eta_1+1}}{\Gamma(\eta_1+1)}) = -\frac{\partial}{\partial \eta_2}^2((\eta_1+1)\log(-\eta_2) - \log(\Gamma(\eta_1+1))) \\ &= -\frac{\partial}{\partial \eta_2}^2((\eta_1+1)\log(-\eta_2)) = (\eta_1+1)\frac{\partial}{\partial \eta_2}\frac{1}{\eta_2} = -\frac{\eta_1+1}{\eta_2^2} = \frac{a}{b^2}\end{aligned}$$

3.

As mentioned in the Bishop (p.101), there is a conjugate prior (i.e. a prior that leads to the posterior distribution having the same functional form as the prior) for every distribution of the exponential family. In case of the Poisson distribution, this is the Gamma distribution.

$$\begin{aligned}p(\lambda|x) &= \frac{p(x|\lambda)p(\lambda)}{p(x)} = \frac{\mathrm{Pois}(x|\lambda)\,\mathrm{Gam}(\lambda|a,b)}{\int_{\lambda_i} \mathrm{Pois}(x|\lambda_i)\,\mathrm{Gam}(\lambda_i|a,b)d\lambda_i} \\ &= \frac{\frac{\lambda^x}{x!}e^{-\lambda}\frac{1}{\Gamma(a)}b^a\lambda^{a-1}e^{-b\lambda}}{\int_{\lambda_i}\frac{\lambda_i^x}{x!}e^{-\lambda_i}\frac{1}{\Gamma(a)}b^a\lambda_i^{a-1}e^{-b\lambda_i}}\end{aligned}$$

$$
\begin{aligned}
&= \frac{\lambda^x e^{-\lambda} \lambda^{a-1} e^{-b\lambda}}{\int_{\lambda_i} \lambda_i^x e^{-\lambda_i} \lambda_i^{a-1} e^{-b\lambda_i}} \\
&= \frac{\lambda^{a-1+x} e^{-b\lambda-\lambda}}{\int_{\lambda_i} \lambda_i^{a-1+x} e^{-b\lambda_i-\lambda_i}} \\
&\propto \lambda^{a-1+x} e^{-b\lambda-\lambda} \\
&\propto \mathrm{Gam}(\lambda | a+x, b+1)
\end{aligned}
$$

As we can see, the resulting posterior is of the same functional form as the prior, i.e. a Gamma distribution.